T. Hirashima et al. (Eds.) (2011). Proceedings of the 19th International Conference on Computers in Education. Chiang Mai, Thailand: Asia-Pacific Society for Computers in Education

Similar Movie Search System by Co-occurrenceWordson VOD Lecture with Japanese Subtitle

Nobuyuki KOBAYASHI[a*], Noboru KOYAMA[b], HiromitsuSHIINA[c] &Fumio KITAGAWA[c]
[a]*Faculty of Human Sciences, Sanyo Gakuen University, Japan*
[b]*GraduateSchool of Informatics, OkayamaUniversity of Science, Japan*
[c]*Faculty of Informatics, OkayamaUniversity of Science, Japan*
*koba_nob@sguc.ac.jp

Abstract: Asearch system for VOD lectures is useful if it beyond the search of only text. To facilitate better searching for movie segments of VOD lectures with Japanese subtitles, we propose a method of using subtitles and a solving maximum likelihood detection from a max mixture of Gaussian distributions. The detection was performed by a statistical method by using the EM algorithm. In addition, we propose similar movie search system which provides a movie segment by co-occurrence word as the detectedsegment in case a word frequency is notenough in subtitle.

Keywords: Movie search, Gaussian mixture model, VOD System, co-occurrence word

Introduction

In recent years, e-learningsystems have proliferated on both the Internet and organization intranets. E-learning systemssuch as video on demand (VOD) andweb-based training systems are used in many universities.In Okayama University of Science, e-learning systems that employ VOD have been used since 2004[1]. Many students are attending lectures that are delivered using the VOD service. However students use VOD service for self-learning after attending lecture, they have to watch from beginning of VOD contents. The topic retrieval function is necessary and indispensable.As methods of detecting movie segments of VOD lectures, we propose a method ofdetecting movie segments by adapting a Gaussian mixture modelforVOD lectures with Japanese subtitles. The parameters of the model are computed by a statistical expectation-maximization (EM)algorithm[1]. This method fits the frequency distribution of a search word with a Gaussian mixture model and detects a movie segment form each Gaussiandistribution (normal distribution). In particular, the characterization of this study is to use the subtitles of a speaker for the estimation of the number of movie segments. However, itis difficult to detectin case there is few search word in frequency of appearance.In this study, as an additional facilitate of movie search system on VOD system, we provide similar movie search system which use co-occurrence words[1]of a search word. Rankingof co-occurrence words is computed by three factors, these are the number of the phrases from a search word, frequency of two words occurrence in same subtitle and frequency of co-occurrence word. Therefore the detected movie segment is provided by a high frequency of co-occurrence word.

1. e-Learning system and search system

The screenshot of VOD system is shown in Figure 1. It displays a movie in the upper left of screen, a list of contents on the lower left and a slide on the right side.One lecture is divided into three sections. Furthermore, homework regarding the contents of the lecture is

displayed at the end of each lecture. It is provided to test the understanding of the lecture. In this study, we have added a new search system based on the histogram of the search word frequency in each interval, as shown on the right side of Figure2.

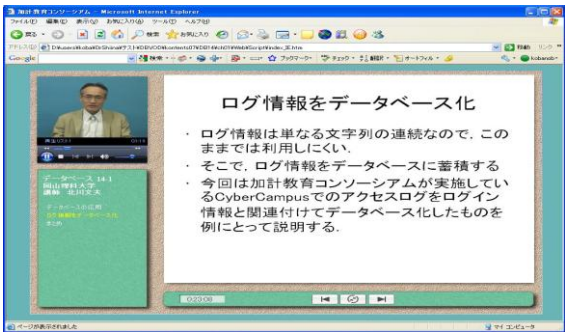

Figure1 Screenshot of lecture screen

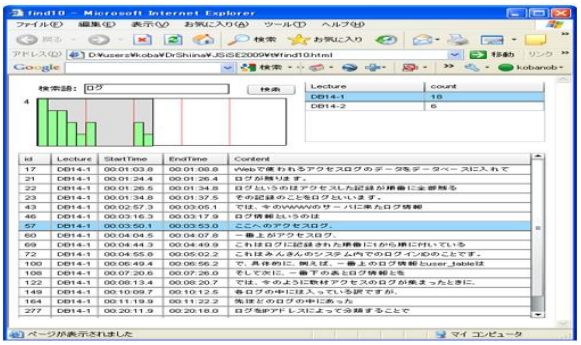

Figure 2 Screenshot of search screen

2. Estimation of movie segments by a mixture of Gaussian distributions

In this study, we suppose that one simple histogram peak corresponds to one topic of a movie segment and the histogram is synthesized by several simple peaks. For example, we use the search word as "広告" in the 14th lecture of the"database" of the cyber campus at Okayama University of Science. The meaning of "広告" is an "advertisement." It uses the keyword in the 14th homework . The graph of the word-frequency histogram and the approximate curve are shown in Figure 3. The horizontal axis of Figure 3 shows the movie time. The histogram has a bin width of 1 min. Thus the vertical axis shows the frequency of words once every minute. As shown in Figure 3, we couldconsider that the histogram has been synthesizedby four distinct peaks. Accordingly, weconsider that the approximate curve is synthesized by amixture of Gaussian distributions from the histogram of a particular search word. The parameter estimation of a Gaussian mixture model is determined using an EM algorithm. Because it is a linear combination of a Gaussian distribution (normal distribution), it is possible to detect a movie segment from the estimated mixture of Gaussian distributions. The definition of the appearance-time of a search word, the Gaussian distribution and the mixture of Gaussian distributions as follows.

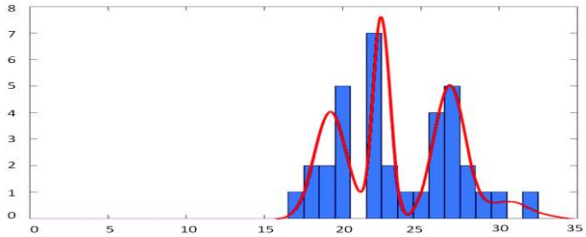

Figure 3 The approximation of a mixture of Gaussian

- Let nbe the number of search words, $X = \{x_1, x_2, \cdots, x_n\}$be a set of appearance-times.
- Gaussian distribution (normal distribution) : $\phi(x; \mu_l, \sigma_l^2) = \frac{1}{2\pi\sigma_l^2} \exp\left(-\frac{(x-\mu_l)^2}{2\sigma_l^2}\right)$.
- Mixture of Gaussian distribution: $q_t(x;\theta) = \sum_{l=1}^{m} w_l \bullet \phi\left(x; \mu_l, \sigma_l^2\right)$.

The number of component Gaussian distributions of Gaussian mixture model is shown $l = 1, \cdots m$, the parameters of the component Gaussian distributionsare $\theta = \{\mu_1, \cdots, \mu_m, \sigma_1^2, \cdots, \sigma_m^2\}$, the weight of the l-th Gaussian distribution isw_l, the average of the l-thGaussian distribution isμ_l, the variance of l-thGaussian distribution isσ_l^2. The algorithm for the estimation of movie segments consists of two processes. In the first

process, the EM algorithm estimates the position of the component Gaussian distributions. The second process is the interval-estimation processing of the beginning and the ending.

(1) EM Algorithm: Parameters of Gaussian mixtures are estimated by the EM algorithm.
- Initial step: Let μ_l be the midpoint whichdivided movie times in l,$\sigma_l = 1$, and $w_l = 1$.
- E-Step: $\hat{\eta}_{i,l} = \dfrac{\hat{w}_l \bullet \phi\left(x_i; \hat{\mu}_l, \hat{\sigma}_l^{\,2}\right)}{\sum_{l'=1}^{m} \hat{w}_l \bullet \phi\left(x_i; \hat{\mu}_{l'}, \hat{\sigma}_{l'}^{\,2}\right)}$
- M-Step: $\hat{w}_l := \dfrac{1}{n}\sum_{i=1}^{n} \hat{\eta}_{i,l}$, $\hat{\mu}_l := \dfrac{\sum_{i=1}^{n} \hat{\eta}_{i,l} \bullet x_i}{\sum_{i'=1}^{n} \hat{\eta}_{i',l}}$, $\hat{\sigma}_l^{\,2} := \dfrac{\sum_{i=1}^{n} \hat{\eta}_{i,l} \bullet \left(x_i - \hat{\mu}_l\right)^2}{\sum_{i'=1}^{n} \hat{\eta}_{i',l}}$.

(2) Estimation of movie segments: The Gaussian mixture model of word frequency distribution which comprises the appearance time for the word is approximated by an EM algorithm. In our VOD lecture, because each section of lecture is anywhere between 20 and 30 min, we assume thatthere are a maximum of five movie segments. These are approximated with the superposition of Gaussian distributions as *m*=1, ..., 5.From each component's Gaussian distribution of the Gaussian mixture model, we detect that the movie segment has a standard deviation of$\pm\sigma_l$width from the average$\mu_l$which covers 68.26%.

3. Ranking a movie segments

The above detect of movie segments are notenough for movie search system, then we rank the detected movie segment by another mixture of Gaussian distribution in addition to the distribution by the Akaike's information criterion (AIC). In this study, we consider evaluation method torankmovie segments. The main point is to evaluate a distribution which removes the *k*-thGaussian distribution from m Gaussian distributions. This distribution is defined by the following formula.

$$q'_{t,k}(x;\theta) = \sum_{l=1, l\neq k}^{m} w_l \bullet \phi\left(x; \mu_l, \sigma_l^{\,2}\right).$$

The evaluation of each movie segment computes the influence of the removed distribution on the whole distribution during the movie time$t = \mu_l - \sigma_l$ .. $t = \mu_l + \sigma_l$. In this study, we propose the following evaluation.

$$V\ (x;\theta) = \frac{1}{2}\left(\sum_{i=1}^{n}\left(q_t(x) \bullet \log\frac{q_t(x)}{q'_{t,k}(x;\theta)}\right) + \sum_{i=\mu-\sigma}^{\mu+\sigma}\left(q'_{t,k}(x) \bullet \log\frac{q'_{t,k}(x)}{q_t(x;\theta)}\right)\right).$$

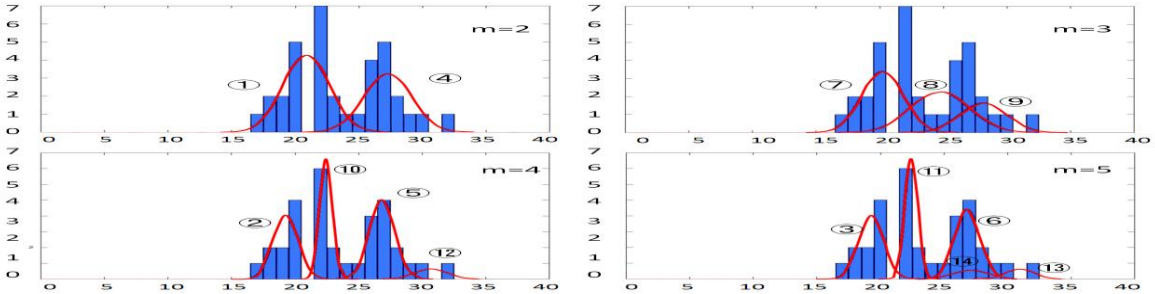

Figure 4Rankings of Gaussian distribution of evaluation by$V\ (x;\theta)$

The characteristic of this evaluation is the average of Kullback-Leibler divergence which is the difference of the two density functions from the mutual function.The example

of the ranking by the evaluation $V(x;\theta)$is shown in Figure4. Each rank by the evaluation of movie segments is shown by the circled numbers.

4. Detection of movie segments using co-occurrence word

The fitting of the Gaussian mixture distribution is difficult in case there is few search word in frequency of appearance.Therefore, it is difficult to detect. As an additional facilitate of movie search system on VOD system, the similar movie search system will be necessary. In this study we provide similar movie segments by using co-occurrence words of a search word. Ranking of co-occurrence words is computed by the number of the phrases from a search word, frequency of two words occurrence in same subtitle and frequency of co-occurrence word. Then the detected movie segment is provided by a high frequency of co-occurrence word. It uses the co-occurrence word as a search word, because it is difficult to make fit Gaussian distributions to the frequency distribution of the search word.We define the ratio of co-occurrence between a word w_1 and a word w_2 by the following formula.

$$Cov(w_1, w_2) = \frac{\sqrt{Cof(w_1, w_2) \bullet feq(w_2)}}{D(w_1, w_2) + 1}$$

$D(w_1, w_2)$: Minimum of phrase distance, $Cof(w_1, w_2)$: Co - occurrence frequency, $feq(w_2)$: Frequency of appearance

Figure5 shows an example of phrase distance in Japanese. When w_1=income(収入) and w_2 =search engine(検索エンジン), the distance of between phrases is two in Japanesesubtitles. An example of sentenceis split into phrases with ChaSen[3] which is Japanese morphological analyzer.

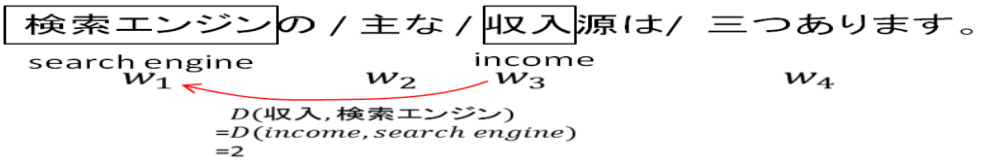

Figure 5phrase distance

4.1 Examples of co-occurrence word

Examples of co-occurrence value for search word “advertisement” (広告) and “income”(収入) is showed in Table 1 and Table 2.

Table 1 Examples of co-occurrenceword of “advertisement” (広告),w_1=advertisement

Rank	Co-occurrence word w_2	$D(w_1,w_2)$	$Cof(w_1,w_2)$	feq(w_2)	$Cov(w_1,w_2)$
1	keyword	0	14	34	21.81
2	Page	1	7	24	6.48
3	search engine	1	5	24	5.48

Table 2 Examples of co-occurrence word of “income” (収入), w_1=income

Rank	Co-occurrence wordw_2	$D(w_1,w_2)$	$Cof(w_1,w_2)$	feq(w_2)	$Cov(w_1,w_2)$
1	search engine	2	4	24	3.27
2	keyword advertisement	2	5	14	2.79
3	use	1	3	7	2.29

4.2 Preliminary Evaluation

We have evaluated this search system from the viewpoint of users. Top five movie segments of search words as "advertisement" and "income"and some co-occurrence words as "keyword" and "search engine" were questioned. The subjects were sixstudents inthe graduate school of informatics, the results of questionnaires are show in Table 3. The evaluation of advertisement as the search word is good, but the evaluation of income is bad. The frequency will have significant impact. In evaluation of similar movie segment, it begins to be divided. In a sense, it is the expected results.

Table 3 The evaluation by user

Question			Answer			
	Search word or co-occurrence word	Ranking of movie segment (start, end)	Strongly Agree	Agree	Disagree	Strongly disagree
Do you satisfied with detected movie section by a search word?	advertisement	① (17:28,24:14)	1	5	0	0
		② (17:25,20:13)	3	2	1	0
		③ (18:17,20:17)	2	3	1	0
		④ (23:42,30:50)	2	4	0	0
		⑤ (25:36,27:40)	2	2	2	0
	income	① (17:26,20:17)	1	5	0	0
		② (24:56,33:22)	0	2	4	0
		③ (28:07,31:47)	0	0	5	1
		④ (28:09,31:45)	0	0	5	1
		⑤ (17:17,19:56)	2	4	0	0
Do you satisfied with detected similar movie section by a search word?	keyword	① (07:02,37:50)	0	1	4	1
		② (03:37,10:58)	0	4	2	0
		③ (06:37,08:02)	1	3	2	0
		④ (06:48,7:47)	1	3	1	1
		⑤ (16:20,35:20)	1	2	3	0
	search engine	① (00:00,05:13)	0	6	0	0
		② (00:00,10:56)	2	3	1	0
		③ (13:23,20:17)	0	5	0	1
		④ (30:35,30:37)	0	0	3	3
		⑤ (30:35,30:36)	0	0	3	3

5. Conclusions and Future Work

In the case of some search words emerge into the subtitle, we consider that it is possible to detected movie segments. On the other hand, in the case of few frequency of a search word, our system provides the similar movie segments by using co-occurring word. A VOD lecture talks about topics by one speaker. Therefore our proposed method by fitting of Gaussian distribution will be effective. As future work, we consider using information of PPTs or Slides for the more precise estimation of movie segments.

References

[1] Dempster, A.P., Laird, N. M., Rubin, D. B. (1977). Maximum likelihood form incomplete data viathe EM algorithm, Journal of the Royal Statistical Society series B, Vol. 39, No.1, pp.1-38.
[2] Kitagawa, F., Onishi S. (2007). An Experiment on selective Course with Face-to-Face or/and E-learning, and Students Behavior(in Japanese),Japan Society of Educational Information, Vol.22 No.3 pp.57-66.
[3] Yamashita,T., Matsuda,H. Matsumoto,Y et al. ChaSen, http://chasen-legacy.sourceforge.jp/
[4] Li, Y. et al. (2006). Sentence Similarity Based on Semantic Nets and Corpus Statistics, IEEE Transactions on Knowledge and Data Engineering, Vol.18, No. 8, pp. 1138-1150.